A Spectral Approach to Item Response Theory

Duc Nguyen
Department of Computer and Information Science
University of Pennsylvania
mdnguyen@seas.upenn.edu

Anderson Y. Zhang
Department of Statistics and Data Science
University of Pennsylvania
ayz@wharton.upenn.edu

Abstract

The Rasch model is one of the most fundamental models in *item response theory* and has wide-ranging applications from education testing to recommendation systems. In a universe with n users and m items, the Rasch model assumes that the binary response $X_{li} \in \{0,1\}$ of a user l with parameter θ_l^* to an item i with parameter β_i^* (e.g., a user likes a movie, a student correctly solves a problem) is distributed as $\mathbb{P}(X_{li}=1) = 1/(1+\exp(-(\theta_l^* - \beta_i^*)))$. In this paper, we propose a *new item estimation* algorithm for this celebrated model (i.e., to estimate β^*). The core of our algorithm is the computation of the stationary distribution of a Markov chain defined on an item-item graph. We complement our algorithmic contributions with finite-sample error guarantees, the first of their kind in the literature, showing that our algorithm is consistent and enjoys favorable optimality properties. We discuss practical modifications to accelerate and robustify the algorithm that practitioners can adopt. Experiments on synthetic and real-life datasets, ranging from small education testing datasets to large recommendation systems datasets show that our algorithm is scalable, accurate, and competitive with the most commonly used methods in the literature.

1 Introduction

Item response theory (IRT) is the study of the relationship between latent characteristics (a student's ability versus a test's difficulty or a user's taste versus a movie's features) and the manifestations of these characteristics (a student's performance on a test or a user's rating of a movie). Originally developed by the psychometric community [45, 51], item response theory has been applied to diverse settings such as education testing [37], crowdsourcing [54], recommendation systems [12], finance [50] and marketing research [10].

One of the most fundamental models in IRT is the Rasch model [45]. It models the *binary response* $X_{li} \in \{0,1\}$ of user l with latent parameter $\theta_l^* \in \mathbb{R}$ to item i with latent parameter $\beta_i^* \in \mathbb{R}$ by

$$\mathbb{P}(X_{li}=1) = \frac{1}{1+\exp(-(\theta_l^* - \beta_i^*))} . \tag{1}$$

For example, in education testing, θ_l^* corresponds to the ability of student l, β_i^* the difficulty of problem i and $X_{li}=1$ if the student correctly solves the problem. Binary response data has grown abundantly in modern domains: Netflix famously switched from a 5-star rating system to a binary like/dislike feedback system, data on students' engagement and performance grows significantly as education moves online during the pandemic.

Traditionally, the goal of estimation under the Rasch model is to *recover the item parameters* β^*. In education testing, an estimate of the item parameters can be used to calibrate scores across different

36th Conference on Neural Information Processing Systems (NeurIPS 2022).

https://doi.org/10.52202/068431-2813

versions of a test. In recommendation systems, the item parameters can be used to produce a ranking over the items. In general, estimation is challenging under the Rasch model because for each user and item pair, we only get a single observation or none in the case of missing data.

Joint maximum likelihood estimate (JMLE) is one of the earliest methods developed for the estimation problem [3, 22, 25, 27]. It estimates both the user and item parameters by maximizing the joint likelihood function using an alternating maximization algorithm. While efficient, JMLE is known to be *inconsistent* (that is, even as $n \to \infty$, JMLE does not recover β^*) when the number of items is finite [3, 24] (e.g., Figure 1a). Intuitively, this is because there are many nuisance user parameters to a finite number of item parameters. As a result, JMLE is mostly used for prelimary parameter estimation and researchers have developed other solutions to address the inconsistency problem, broadly consisting of 3 approaches as follows.

The first approach is marginal maximum likelihood estimate (MMLE) [7]. The statistician first specifies a prior distribution over the user parameters. The objective of MMLE is to maximize the marignal likelihood function which *integrates out* the user parameters. In pratice, MMLE runs quite fast, handles missing data well and is reasonably accurate. However, its performance *depends on the accuracy of the prior distribution*. If misspecified, MMLE may produce inaccurate estimates (e.g., Figure 1b). Model selection is thus a crucial procedure when applying MMLE to real data.

The second approach is conditional maximum likelihood estimate (CMLE) [3, 22, 27]. CMLE builds on the fact that under the Rasch model the total number of positive responses s_l for each user l is a sufficient statistic for the user parameter θ_l^*. Instead of the joint likelihood function, CMLE maximizes the likelihood conditioned on $\{s_l\}_{l=1}^n$. Unlike JMLE, CMLE is *statistically consistent without requiring any distribution assumptions about* θ^*. For small datasets with no missing data, CMLE is quite accurate. However, it may incur *high computational cost and numerical issues* on large datasets with many items and missing entries. Practioners have observed that CMLE often produces inaccurate estimates [35, 36] in this regime (e.g., Figure 1c).

The third approach, which our algorithm follows, uses *pairwise information* that can be extracted from binary responses. Intuitively, if a user responds to two items, one negatively and one positively, we learn that the later is 'better'. Following this intuition, previous authors [23, 17, 48] have designed spectral algorithms that first construct an item-item matrix and then compute its leading eigenvector. One common limitation of these methods is that the item-item matrix is *assumed to be dense*. Therefore, these methods aren't directly extendable to large scale datasets in applications such as recommendation systems where the item-item observation is sparse.

Furthermore, most theoretical guarantees for the above methods are asymptotic $(n \to \infty)$. However, having finite sample error guarantees is useful in real-life applications. For example, when we only observe a handful of responses to a new item, it is important to have an accurate estimate of the error over the item parameter. Asymptotic guarantees, on the other hand, are accurate mostly under a large sample size regime, and can be inaccurate in the data-poor regime.

Our Contributions: Motivated by known limitations of the existing methods, we propose a new, theoretically grounded algorithm that addresses these limitations and performs competitively with the most commonly used methods in the literature. More specifically:

- In Sections 2 and 4, we describe the spectral algorithm and practical modifications – an accelerated version of the original algorithm and a regularization strategy – that allow the algorithm to scale to large real-life datasets with sparse observation patterns and alleviate numerical issues.
- In Section 3, we present *non-asymptotic* error guarantees for the spectral method – the first of their kind in the literature – in Theorems 3.1 and 3.3. Notably, under the regime where m grows, the spectral algorithm has optimal (up to a constant factor) estimation error achievable by any unbiased estimator (Theorem 3.4). Under the challenging regime where m is a constant or grows very slowly we show that the spectral algorithm is, unlike JMLE, *consistent* (Corollary 3.2).
- In Section 5, we present experiment results on a wide range of datasets, both synthetic and real, to show that our spectral algorithm is *competitive* with the most commonly used methods in the literature, *works off-the-shelf with minimal tuning and is scalable* on large datasets.

https://doi.org/10.52202/068431-2813

1.1 Notations and Problem Formulation

As briefly described before, in a universe of n users and m items, each user l has a latent parameter $\theta^*_l \in \mathbb{R}$ and each item i has latent parameter $\beta^*_i \in \mathbb{R}$. The reader may recognize that there is a fundamental identifiability issue associated with the Rasch model pertaining to translation. That is, $\{\theta^*, \beta^*\}$ and $\{\theta^* + \alpha \mathbf{1}_n, \beta^* + \alpha \mathbf{1}_m\}$ describe the same model for any $\alpha \in \mathbb{R}$. For this reason, we impose a normalization constraint on the item parameters ${\beta^*}^\top \mathbf{1}_m = 0$. We consider the fixed range setting where $\beta^*_i \in [\beta^*_{\min}, \beta^*_{\max}]\, \forall i \in [m]$ for some constants $\beta^*_{\min}$, $\beta^*_{\max}$. Similarly, we assume that $\theta^*_l \in [\theta^*_{\min}, \theta^*_{\max}]$ for some constants $\theta^*_{\min}, \theta^*_{\max}$ [1]. The observed data is $X \in \{0, 1, *\}^{n \times m}$ where $*$ denotes missing data and for entries where $X_{li} \neq *$, X_{li} is independently distributed per Equation (1). Let $A \in \{0,1\}^{n\times m}$ denote the assignment matrix where $A_{li} = 1$ if user l responds (either negatively or positively) to item i and 0 if user l does not respond to item i (i.e., $X_{li} = *$). Define $B = A^\top A$, i.e., B_{ij} is the number of users who respond to both items i, j. The goal of item estimation is to obtain an estimate β from the observed data X and the metric of interest is the ℓ_2 error, $\|\beta - \beta^*\|_2$.

2 The Spectral Estimator

In this section we describe our spectral algorithm which is summarized in Algorithm 1. At a high level, the algorithm constructs a Markov chain defined on a graph whose vertices are the items and its transition probabilities are estimated using the observed user-item response data. The algorithm then computes the stationary distribution of this Markov chain and the estimate β is obtained following a simple transformation.

We first define, for each item pair i, j and a fixed assignment A, a quantity which we term *pairwise differential measurement*:

$$Y_{ij} = \sum_{l=1}^{n} A_{li} A_{lj} X_{li}(1 - X_{lj}) \quad \forall i \neq j \in [m]\,. \tag{2}$$

Intuitively, Y_{ij} is the number of users who respond 1 to i and 0 to j. Given the pairwise differential measurements, consider a Markov chain $P \in [0,1]^{m\times m}$ whose transition probabilities are defined as follows:

$$P_{ij} = \begin{cases} \frac{1}{d} Y_{ij} & \text{if } i \neq j \\ 1 - \sum_{k \neq i} \frac{1}{d} Y_{ik} & \text{if } i = j \end{cases} \,, \tag{3}$$

where d is a sufficiently large normalization factor chosen such that the resulting pairwise transition probability matrix does not contain any negative entries. Typically, $d = O(\max_{i\in[m]} \sum_{k\neq i} B_{ik})$. The algorithm then computes the stationary distribution π of the Markov chain (e.g., using power iteration) and recover β using a post-processing step. In real-life datasets, the constructed Markov chain is often sparse (not every pair of items has non-zero pairwise differential measurements). Practicioners could take advantage of this sparsity to speed up the computation of the stationary distribution such as by using sparse matrix-vector multiplication subroutines.

To understand the intuition behind our spectral algorithm, let us consider the following idealized Markov chain where the state transition probabilities are exact:

$$P^*_{ij} = \begin{cases} \frac{1}{d} Y^*_{ij} & \text{for } i \neq j \\ 1 - \frac{1}{d}\sum_{k \neq i} Y^*_{ik} & \text{for } i = j \end{cases} \,, \tag{4}$$

where $Y^*_{ij} = \sum_{l=1}^{n} A_{li} A_{lj} \mathbb{E}[X_{li}(1 - X_{lj})]$. For every pair i, j, given a sufficiently large number of users who respond to both items, Y_{ij} will concentrate around Y^*_{ij}. Then, under an appropriately large scaling factor d, $P_{ij} \approx P^*_{ij}$ and the two Markov chains are 'close'. This means that the stationary distribution of P is also close to that of P^*. At the same time, the true item parameter β^* is directly related to the stationary distribution of P^*. This relation is summarized by Proposition 2.1.

Proposition 2.1. *Consider the idealized Markov chain described in Equation (4). The stationary distribution π^* of P^* satisfies $\pi^*_i = e^{\beta^*_i} / (\sum_{k=1}^{m} e^{\beta^*_k})$ for $i \in [m]$.*

[1] The bounded range assumption is a common one in the literature on the Rasch model. Intuitively, it eliminates the presence of items that are always repsonded positively to (or negatively to) and users who only responds positively (or negatively) that leads to parameter unidentifiability [26].

https://doi.org/10.52202/068431-2813

Algorithm 1 Spectral Estimator

```
Input: User-item binary response data $X \in \{0, 1, *\}^{n \times m}$
Output: An estimate of the item parameters $\beta = [\beta_1, \ldots, \beta_m]$

Construct a Markov chain $P$ per Equation (3)
Compute the stationary distribution of $P$:
    Initialize $\pi^{(0)} = [\frac{1}{m}, \ldots, \frac{1}{m}]$
    For $t = 1, 2, \ldots$ until convergence, compute
        $\pi^{(t)^\top} = \frac{\pi^{(t-1)^\top} P}{\|\pi^{(t-1)^\top} P\|_1}$
Compute $z = \pi / d$ and $\bar{\beta} = \log(z)$
Return the normalized item parameters, i.e., $\beta = \bar{\beta} - \bar{\beta}^\top \mathbf{1}/m$
```

Essentially Proposition 2.1 states that π^* is proportional to e^{β^*}. Thus β^* can be recovered from π^* up to a global normalization. Now, given a sufficiently large number of users, the empirical stationary distribution π will be close to π^* and naturally the obtained estimate β is also close to β^*.

Readers who are familiar with the ranking from pairwise comparison literature might recognize the similarity between the spectral algorithm and Rank Centrality [43] for parameter estimation under the Bradley-Terry-Luce model [38]. Similarly to Rank Centrality, our algorithm constructs a Markov chain on the item-item graph and recovers parameter estimate from its stationary distirbution. In both cases, the Markov chain interpretation is motivated by the unique characteristics of the BTL and Rasch likelihood function. However, the specific construction differs between our algorithm and Rank Centrality and so does the resulting analysis.

3 Theoretical Analysis

In this section, we present the main theoretical contributions of the paper. Specifically, we obtain in Section 3.1 two finite sample error bounds for two different regimes of m: where m is a constant or grows very slowly and where m grows at least logarithmically relative to n. In addition to our upper bounds, we show in Section 3.2 a Cramer-Rao lower bound for the mean squared error of any unbiased estimator, establishing the optimality of the spectral algorithm under the the second regime. For the special case $m = 2$, we show that the error rate obtained by the spectral algorithm is optimal up to a $\log$ factor.

3.1 Finite Sample Error Guarantees

Sampling Model: Let us consider a random sampling model where for each user $l \in [n]$, each item $i \in [m]$ is independently shown to that user with probability p (i.e., $\mathbb{P}(A_{li} = 1) = p$). Once shown an item l, the user i responds with X_{li} distributed according to Equation (1).

Under this sampling model and the regime where m *is a constant or grows very slowly*, we obtain the following upper bound on the estimation error of the spectral algorithm which is, to the best of our knowledge, the first finite sample error guarantee for any consistent estimator under the Rasch model in the literature.

Theorem 3.1. *Consider the sampling model described above. Suppose that $np^2 \geq C' \log m$ for a sufficiently large constant C' then the output of the spectral algorithm statisfies*

$$\|\beta - \beta^*\|_2 \leq \frac{C\sqrt{\max\{m, \log np^2\}}}{\sqrt{np^2}}$$

with probability at least $1 - \min\{e^{-12m}, \frac{1}{(np^2)^{12}}\} - \exp\left(-C_1 np^2\right)$, where C, C_1 are constants.

As alluded to before in our algorithm description, the proof of Theorem 3.1 uses Markov chain analysis and a central object is the idealized Markov chain P^* with its stationary distribution π^*. The proof is rather long and involved so we defer the details to the supplementary materials and describe here the main idea. The starting point is a Markov chain eigen-perturbation bound (see Lemma A.3):

$$\|\pi - \pi^*\|_2 \leq \frac{\|{\pi^*}^\top (P^* - P)\|_2}{\mu^*(P^*) - \|P - P^*\|_2}\,,$$

where $\mu^*(P^*)$ is the *spectral gap* of the idealized Markov chain. We then bound the numerator and the denominator separately. We will show under the setting of Theorem 3.1 that

$$\mu^*(P^*) - \|P - P^*\|_2 = \Omega\bigg(\frac{1}{d}\bigg) \quad \text{and} \quad \|{\pi^*}^\top (P^* - P)\|_2 = O\bigg(\frac{\sqrt{\max\{m, \log np^2\}}}{dm\sqrt{np^2}}\bigg)\,.$$

Combining these bounds with the following relation gives us the desired error bound:

$$\|\beta - \beta^*\|_2 = O\bigg(m \cdot \|\pi - \pi^*\|_2\bigg)\,.$$

As an immediate consequence of Theorem 3.1, we can also prove the consistency of the spectral algorithm under the constant m regime. As mentioned previously, JMLE, one of the most well known methods in the Rasch modeling literature, is inconsistent in this regime.

Corollary 3.2. *Consider the setting of Theorem 3.1. For a fixed m and $p = 1$, the spectral algorithm is a consistent estimator of β^*. That is, its output β satisfies* $\lim_{n\to\infty} \mathbb{P}(\|\beta - \beta^*\|_2 < \epsilon) = 1\,, \forall \epsilon > 0\,.$

Under the regime where m grows, we could sharpen the results of Theorem 3.1. Specifically, when the number of items shown to each user is sufficiently large, we improve by a $\sqrt{p}$ factor which can be significant when p is small. This is summarized by the following theorem.

Theorem 3.3. *Consider the setting of Theorem 3.1. Assume further that $mp \geq C'' \log n$ for a sufficiently large constant C'' then the output of the spectral algorithm statisfies*

$$\|\beta - \beta^*\|_2 \leq \frac{C^* \sqrt{m}}{\sqrt{np}}$$

with probability at least $1 - \exp\left(-C_2 np^2\right) - 2n^{-9}$, where C^*, C_2 are constants.

The reader may wonder why there would be a difference between the two regimes. Intuitively, when m is a small constant, the distribution of the number of items shown to the users are not tightly concentrated. Some users are shown all of the items while some are shown only one. By design, our spectral algorithm uses pairwise differential measurements. This means that when a user responds to only one item, that information is not fully used. On the other hand, when $mp = O(\log n)$, the number of items shown to the users is concentrated (all users are shown approximately the same number of items) and more pairwise differential measurements are available. There is less information being under-utilized by the algorithm and it enjoys a tighter (in fact optimal) error rate.

3.2 Cramer-Rao Lower Bound

In this section, we present complementary results to our finite error guarantees obtained in the previous section. Notably, under the regime where m is allowed to grow with n, we show that the minimum mean squared error achievable by any unbiased estimator is *no more than a constant factor* smaller than the upper bound for the spectral algorithm established in Theorem 3.3. This optimality result is summarized by the following theorem.

Theorem 3.4. *Consider the sampling model described in in Section 3.1. Let T be any unbiased estimator for the item parameters. Then the mean squared error of such estimator is lower bounded as*

$$\mathbb{E}\|T(X) - \beta^*\|_2^2 \geq \frac{cm}{np}\,,$$

where $T(X)$ is the output of the estimator T when given data X and c is a constant.

Now note that under the settings of Theorem 3.3, the output of the spectral algorithm also statisfies $\|\beta - \beta^*\|_2^2 = O(\frac{m}{np})$. To the best of our knowledge, this is the first non-asymptotic optimality result for any item estimation method under the Rasch model.

As noted before, when the number of items is constant, our error bound in Theorem 3.1 incurs an additional $1/\sqrt{p}$ factor. We now argue that the upper bound obtained there may already be optimal in this challenging regime. Consider the special case when $m = 2$. Essentially, the goal is to estimate the *difference of the two item parameters*. For a particular user l, suppose that we have no information about her parameter θ_l other than the bounded condition. If the user's responses consist of a single response (the response to one item is not observed) or that her responses to both items are identical (either both 0 or 1), then we learn little about the difference between the two items. We refer to these responses as 'bad' responses. The relative difference between the items is only revealed if the user responds differently to the items. As noted in the description of our spectral algorithm, we refer to such information as *pairwise differential measurements*.

For the special case $m = 2$, both CMLE and JMLE actually ignore 'bad' responses. This is because in both algorithms, it has been shown that including bad responses in the respective objective likelihood function leads to parameter unidentifiability [26]. As mentioned earlier, MMLE, requires an accurate prior distribution in order to obtain good estimate accuracy. This is not possible when we have no information about the user parameters. With the exception of MMLE, *all estimation methods* that we are aware of in the literature only use pairwise differential measurements.

With these points considered, if we restrict our attention to the class of algorithms that use pairwise differential measurements, then the spectral algorithm indeed achieves the best possible (up to a log factor) estimation error.

Theorem 3.5. *Fix $m = 2$ and consider the sampling model described in Section 3.1. Let T be any unbiased estimator for the item parameters that only uses* pairwise differential measurements*. Then the mean squared error of such estimator is lower bounded as*

$$\mathbb{E}\|T(X) - \beta^*\|_2^2 \geq \frac{c'}{np^2}\,,$$

where $T(X)$ is the output of the estimator T when given data X and c' is a constant.

As seen from Theorem 3.1, the estimate produced by the spectral algorithm satisfies $\|\beta - \beta^*\|_2^2 = \tilde{O}(\frac{1}{np^2})$, establishing its near optimality.

4 Practical Implementation Aspects

In this section, we discuss two important practical aspects that practicioners may consider when applying the spectral algorithm real-life datasets of which observation pattern may not correspond exactly to the sampling model described in Section 3.1. Firstly, we identify slow convergence as a problem encountered by the spectral algorithm when the data is skewed in the sense that some items are highly responded to by users while some items are rarely responded to. To address this issue, we propose an accelerated spectral algorithm that enjoys the same error guarantees as the original spectral algorithm but runs significantly faster in practice and suffers from fewer numerical issues. Secondly, we discuss regularization strategy when the spectral algorithm is applied to datasets with sparse observation patterns.

Accelerating the Spectral Algorithm: Recall that in the original spectral Algorithm 1, we use a common normalization constant d that generally scales as $O(\max_i \sum_{k\neq i} B_{ik})$. In practice, the distribution of the user-item assignment could be skewed such that some items are rarely responded to while some elicit many user responses. In such cases, the items with few responses will have few pairwise differential measurements $\ll d$. The induced Markov chain will contain large self-loops for these items. We observe that these large self-loops lead to a slower convergence when computing the stationary distribution and more numerical issues. This observation was also noted in [2] and we propose a similar solution to eliminate large self-loops that is to use a *different normalizing factor for each vertex*. Consider the following modified Markov chain:

$$\bar{P}_{ij} = \begin{cases} \frac{1}{d_i} Y_{ij} & \text{if } i \neq j \\ 1 - \frac{1}{d_i}\sum_{k\neq i} Y_{ik} & \text{if } i = j \end{cases}\,, \tag{5}$$

where Y_{ij} is defined in Equation (2) and $\{d_i\}_{i=1}^m$ are appropriately chosen normalization factors such that the resulting transition probability matrix does not contain any negative entries. In our experiments, we choose $d_i = O(\sum_{k \neq i} B_{ik})$. The accelerated spectral algorithm computes the stationary distribution of the above modified Markov chain, and recovers the item parameters via a post-processing step. The algorithm is summarized in Algorithm 2.

Algorithm 2 *Accelerated* Spectral Estimator

```
    Input: User-item binary response data $X \in \{0, 1, *\}^{n \times m}$
    Output: An estimate of the item parameters $\beta = [\beta_1, \ldots, \beta_m]$
Construct a modified Markov chain $\bar{P}$ per Equation (5)
Compute the stationary distribution $\bar{\pi}$ of $\bar{P}$
Compute $z = D^{-1}\bar{\pi}$ and $\bar{\beta} = \log(z)$ where $D = \mathrm{diag}(d_1, \ldots, d_m)$
Return the normalized item parameters, i.e., $\beta = \bar{\beta} - \bar{\beta}^\top \mathbf{1}/m$
```

Interestingly, the accelerated algorithm produces essentially the same estimate as the original algorithm. Under some regularity conditions, there is a direct one-to-one relation between the stationary distribution π obtained using the original Markov chain in Algorithm 1 and the stationary distribution $\bar{\pi}$ of the Markov chain parametrized by $\bar{P}$. This result is summarized in Theorem 4.1.

Theorem 4.1. *Consider the modified Markov chain $\bar{P}$ constructed per Equation (5) and the original Markov chain P constructed per Equation (3). Suppose that both $\bar{P}$ and P admit unique stationary distributions $\bar{\pi}$ and π, respectively. Then*

$$\bar{\pi}_i = \frac{\pi_i d_i}{\sum_{k=1}^m \pi_k d_k} \quad \forall i \in [m] \,,$$

where d_i are the normalization factors in the construction of the modified Markov chain $\bar{P}$.

Assuming *perfect numerical precision*, the two versions of the spectral algorithm output the same estimates. Therefore the guarantees in Theorems 3.1 and 3.3 also apply to the accelerated spectral algorithm. However, in our experiments, we observe that the accelerated spectral algorithm converges much faster and suffers from fewer numerical issues than the original version on real-life datasets, leading to better performance overall. We thus use the accelerated version in our experiments.

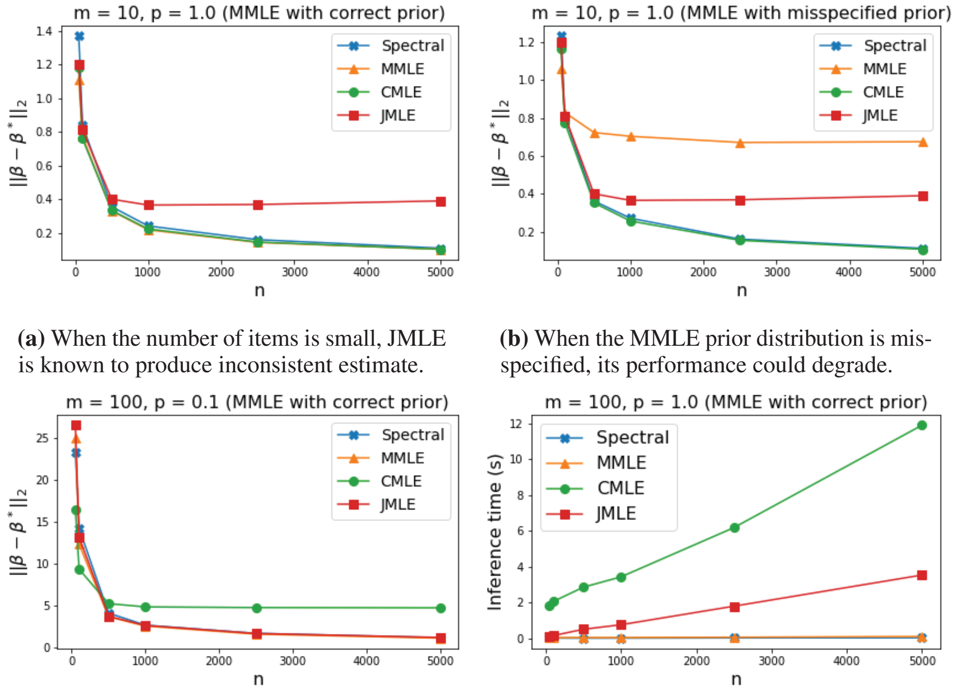
Regularization for Sparse Datasets: In some real-life datasets, we observe that certain pairs of items have few pairwise differential measurements. Furthermore, the pairwise differential data is one-sided (e.g., users who respond to the two items always respond positively to one but negatively to the other). This could happen to pairs that have been shown to only few users. The existence of many such pairs may also introduce numerical issues and parameters unidentifiability. For example, when there is an item i such that $Y_{ji} = 0 \, \forall j \neq i$, the stationary probability correpsonding to this item will be 0 and the item parameter estimate will be $-\infty$. As a solution, we propose adding regularization in the construction of the Markov chain. Specifically, for every pair of items i, j such that $B_{ij} > 0$ redefine

$$Y_{ij} = \sum_{l=1}^{n} A_{li} A_{lj} X_{li} (1 - X_{lj}) + \nu \quad ,$$

where ν is a small constant. In all of our experiments on real-life datasets, we use $\nu = 1$ and find that the regularization parameter requires little tuning. Regularization also ensures the uniqueness of the stationary distribution. So long as the graph underlying the Markov chain is connected, adding regularization ensures that no pairwise transition probability is 0. This makes the constructed Markov chain *ergodic* and there is a unique stationary distribution [44].

5 Experiments

In this section, we present empirical findings which support the practical value of our spectral algorithm. Our baselines are the most commonly used estimation algorithms in the literature: conditional maximum likelihood estimate (CMLE), marginal maximum likelihood estimate (MMLE) and joint maximum marginal likelihood (JMLE). Their open source implementation can be found online [49]. We include the python implementation of our spectral algorithm in the supplementary materials.



(a) When the number of items is small, JMLE is known to produce inconsistent estimate.

(b) When the MMLE prior distribution is misspecified, its performance could degrade.

(c) In larger datasets with missing data, CMLE's performance could degrade.

(d) The spectral method is comparatively scalable.

Figure 1: *(Synthetic Data Experiments.)* The spectral method performs consistently well (both in terms of ℓ_2 error and time complexity) across a range of settings while CMLE, MMLE and JMLE could underperform in unfavorable settings. Presented results have been averaged over 100 trials.

Synthetic Data: We generate the item parameters β^* from a standard normal distribution and user parameters θ^* from $\mathcal{N}(0, \sigma^2)$ where σ^2 varies for different model settings. Recall that MMLE requires the statistician to specify the prior distribution over the user parameters and for synthetic experiments, we specify this prior distribution to be the standard normal distribution. Subfigures (a)-(c) of Figure 1 show $\|\beta - \beta^*\|_2$ against n for all 4 algorithms under 3 *different model settings* while Subfigure (d) shows inference time. When there is no missing data and MMLE's prior distribution is correctly specified ($\sigma = 1$), CMLE, MMLE and the spectral algorithm perform equally well. However, when MMLE's prior distribution is misspecified ($\sigma = 2$), it produces inconsistent estimates. As mentioned before, JMLE is known to produce inconsistent estimate when m is small relative to n. On the other hand, CMLE's performance degrades under moderately sized dataset with missing data. The spectral algorithm, however, consistently performs well across all of these settings and is comparatively scalable.

Real Data: We perform experiments on a wide range of real-life datasets from education testing datasets to book and movie ratings datasets for recommendation systems. In order to transform ratings data to binary response data, we follow the procedures in previous works [33, 19]. Specifically, for each user, we convert all ratings higher than the average to 0 and 1 otherwise (so that items with a higher parameter value is 'better'). Since real-life datasets do not come with true β^* parameters, we compare the algorithms on four metrics: area under the ROC curve (denoted AUC) on heldout test data; log-likelihood on heldout test data; inference time and top-K accuracy where the reference top-K set is determined by average ratings. We have also followed the standard procedure in the recommendation systems literature to remove items that have few (≤ 10) ratings from the reference top-K set. In our experiments, we consider $K \in \{10, 25, 50\}$. We defer extra experiment results where we include additional algorithms such as Bayesian estimation [42] and pairwise likelihood estimation to the supplementary materials. We mention here a few notable datasets: RIIID [1] ($m = 6k, n = 23k$, education testing dataset), ML-20M [28] ($m = 27k, n = 138k$), Book-Genome [31] ($m = 10k, n = 350k$).

	AUC				Log-likelihood				Top-[10 // 25 // 50] accuracy				Total Inference time (seconds)			
Dataset	Spectral	MMLE	CMLE	JMLE	Spectral	MMLE	CMLE	JMLE	Spectral	MMLE	CMLE	JMLE	Spectral	MMLE	CMLE	JMLE
LSAT	0.707	0.707	0.707	0.707	−0.487	−0.489	−0.487	−0.485	N/A	N/A	N/A	N/A	0.028	0.159	0.154	0.075
UCI	0.565	0.565	0.565	0.565	−0.687	−0.686	−0.692	−0.706	N/A	N/A	N/A	N/A	0.015	0.133	0.136	0.034
3 GRADES	0.532	0.532	0.532	0.532	−0.706	−0.692	−0.699	−0.717	N/A	N/A	N/A	N/A	0.021	0.181	0.105	0.009
RIIID	0.723	0.724	N/A	0.724	−0.486	−0.49	N/A	−0.486	N/A	N/A	N/A	N/A	13.1	104	N/A	61.2
HETREC	0.729	0.729	0.506	0.73	−0.604	−0.603	−1.119	−0.602	0.5 // 0.64 // 0.6	0.0 // 0.0 // 0.02	0.0 // 0.0 // 0.0	0.0 // 0.0 // 0.02	50.1	140	224k	144
ML-100K	0.662	0.659	0.498	0.665	−0.646	−0.66	−1.159	−0.653	0.4 // 0.6 // 0.54	0.0 // 0.0 // 0.0	0.0 // 0.0 // 0.0	0.0 // 0.0 // 0.0	1.39	16.2	9.56k	21
ML-1M	0.698	0.701	0.468	0.7	−0.626	−0.632	−1.166	−0.63	0.8 // 0.72 // 0.72	0.6 // 0.6 // 0.62	0.0 // 0.0 // 0.0	0.5 // 0.64 // 0.66	19.2	86.9	156k	194
EACH MOVIE	0.716	0.718	0.522	0.716	−0.615	−0.613	−0.946	−0.614	0.8 // 0.76 // 0.82	0.8 // 0.68 // 0.84	0.0 // 0.0 // 0.02	0.6 // 0.6 // 0.72	11.3	329	220k	1.9k
ML-10M	0.714	0.716	N/A	0.716	−0.617	−0.619	N/A	−0.618	0.5 // 0.84 // 0.7	0.1 // 0.28 // 0.32	N/A	0.0 // 0.32 // 0.36	821	3.93k	N/A	6.55k
ML-20M	0.72	0.71	N/A	0.71	−0.619	−0.619	N/A	−0.619	0.5 // 0.8 // 0.64	0.3 // 0.44 // 0.4	N/A	0.1 // 0.4 // 0.4	1.58k	5.36k	N/A	4.42k
BX	0.546	0.577	0.503	0.57	−0.618	−0.612	−0.8	−0.617	0.3 // 0.16 // 0.16	0.3 // 0.24 // 0.2	0.0 // 0.0 // 0.02	0.3 // 0.2 // 0.18	205	2.02k	156k	481
BOOK-GENOME	0.658	0.665	N/A	0.654	−0.651	−0.645	N/A	−0.651	0.6 // 0.44 // 0.42	0.3 // 0.32 // 0.34	N/A	0.2 // 0.24 // 0.38	2.53k	2.56k	N/A	4.34k

Table 1: *(Real Data Experiments.)* The spectral method (1st column under each metric) is competitive with the baselines (best results are shaded) especially in terms of ranking metrics [2]. The spectral method is generally the fastest method on large datasets. It *works off-the-shelf with minimal tuning* and is *comparatively accurate*.

Table 1 summarizes the performance of the four methods in our experiments. Note that the first four datasets are education testing datasets and thus there are no top-K ranking metrics being measured. For tuning the prior distribution for MMLE, we select the prior distribution that admits the highest log-likelihood on a validation set. On the other hand, CMLE, JMLE and the spectral method requires minimal model tuning. For small scale education datasets (LSAT, UCI, 3GRADES), there are no missing responses and all methods perform very similarly to one another. For large scale ratings datasets, the spectral method remains competitive with the baselines in terms of AUC and log-likelihood but tends to outperforms in top-K accuracy and is significantly more efficient.

In large scale ratings datasets, we observe that the competitor methods tend to assign large parameter value to items that receive only a few but very high ratings (so after data processing, all of the responses to such items are 0) and these items are not included in the reference top-K set. The spectral method, because it operates on pairwise differentials, is less susceptible to these noisy responses and thus more accurately recovers the items in the reference top-K set.

6 Related Works

The Rasch modeling literature is quite broad and we refer the interested reader to recent surveys [6, 46]. Since its original formulation to model psychological tests outcome [45], the model has been extended to account for more complicated response patterns such as numerical ratings and ordinal responses [4, 5, 55, 39]. Higher-order models that incorporates bias variables such as the 2PL and 3PL model [9] and multivariate models [21] remain active areas of research where machine learning techniques have recently been applied with substantial success [8]. The Rasch inference problem is also closely connected to the 1-bit matrix completion problem in machine learning where we observe a sparse $n \times m$ binary matrix with underlying entrywise probability $f(M)$ where f is a mapping function (e.g., logistic) and M is a real-valued matrix. There, the goal is obtain an estimate of M. A commonly proposed approaches for 1-bit matrix completion based on alternating optimization is exactly joint maximum likelihood estimate [15, 14].

7 Ethical Considerations

Our work proposes an algorithm of which real-life applications very often involve actual human data with sensitive information. For example, the Rasch model is often studied in the context of education testing and psychological testing where the subjects of studies are students and patients. Therefore, deploying our algorithm (or any algorithms in this context) needs to be accompanied by thoughtful and thorough ethical considerations. In this work, we provide the algorithmic tool that lays the foundation for our algorithm and its theoretical guarantee. We believe that a socially

[2] As noted before, CMLE has been observed to underperform on large datasets with missing data. In some of our experiments, CMLE fails due to numerical errors or does not converge. The results for CMLE are marked 'N/A' for these experiments.

constructive application of our algorithm should always be accompanied by detailed explanation of its fundamental limitations, assumptions and decision makers need to take into account these aspects when interpreting the results returned by the algorithm.

8 Conclusion

We propose a new spectral algorithm for the item estimation problem under the celebrated Rasch model. Our algorithm is theoretically well-founded, practically performant and should be added to the statistician's quiver of estimation methods when analyzing binary response data. Extending our algorithm to more expressive IRT models such as 2PL or 3PL and response types is an open avenue. In the future, we also hope to generalize the method to more complicated response data types such as ordinal or rating data, as well as incorporating ancillary information (user and item features).

9 Acknowledgement

The authors would like to thank William Zhang for proofreading and leaving helpful comments on earlier drafts of this paper. We would also like to thank the anonymous reviewers for their thoughtful suggestions that have been incorporated to improve this paper. D.N. and A.Z. are supported by NSF Grant DMS-2112988. A.Z. acknowledges financial support from the Alfred H. Williams Faculty Scholar award. Any opinions expressed in this paper are those of the author and do not necessarily reflect the views of the National Science Foundation.

References

[1] Riiid answer correctness prediction. `https://www.kaggle.com/c/riiid-test-answer-prediction/`. Accessed: 2022-05-11.

[2] Arpit Agarwal, Prathamesh Patil, and Shivani Agarwal. Accelerated spectral ranking. In *International Conference on Machine Learning*, pages 70–79. PMLR, 2018.

[3] Erling B Andersen. Conditional inference for multiple-choice questionnaires. *British Journal of Mathematical and Statistical Psychology*, 26(1):31–44, 1973.

[4] Erling B Andersen. Sufficient statistics and latent trait models. *Psychometrika*, 42(1):69–81, 1977.

[5] David Andrich. A rating formulation for ordered response categories. *Psychometrika*, 43(4):561–573, 1978.

[6] Vahid Aryadoust, Li Ying Ng, and Hiroki Sayama. A comprehensive review of rasch measurement in language assessment: Recommendations and guidelines for research. *Language Testing*, 38(1):6–40, 2021.

[7] Debabrata Basu. On the elimination of nuisance parameters. In *Selected Works of Debabrata Basu*, pages 279–290. Springer, 2011.

[8] Yoav Bergner, Stefan Droschler, Gerd Kortemeyer, Saif Rayyan, Daniel Seaton, and David E Pritchard. Model-based collaborative filtering analysis of student response data: Machine-learning item response theory. *International Educational Data Mining Society*, 2012.

[9] A Lord Birnbaum. Some latent trait models and their use in inferring an examinee's ability. *Statistical theories of mental test scores*, 1968.

[10] Justyna Brzezińska et al. Latent variable modelling and item response theory analyses in marketing research. *Folia Oeconomica Stetinensia*, 16(2):163–174, 2016.

[11] Iván Cantador, Peter Brusilovsky, and Tsvi Kuflik. 2nd workshop on information heterogeneity and fusion in recommender systems (hetrec 2011). In *Proceedings of the 5th ACM conference on Recommender systems*, RecSys 2011, New York, NY, USA, 2011. ACM.

[12] Chih-Ming Chen, Hahn-Ming Lee, and Ya-Hui Chen. Personalized e-learning system using item response theory. *Computers & Education*, 44(3):237–255, 2005.

https://doi.org/10.52202/068431-2813

[13] Pinhan Chen, Chao Gao, and Anderson Y Zhang. Partial recovery for top-k ranking: Optimality of mle and sub-optimality of spectral method. *arXiv preprint arXiv:2006.16485*, 2020.

[14] Yunxiao Chen, Chengcheng Li, and Gongjun Xu. A note on statistical inference for noisy incomplete 1-bit matrix. *arXiv preprint arXiv:2105.01769*, 2021.

[15] Yunxiao Chen, Xiaoou Li, and Siliang Zhang. Joint maximum likelihood estimation for high-dimensional exploratory item factor analysis. *Psychometrika*, 84(1):124–146, 2019.

[16] Yuxin Chen, Jianqing Fan, Cong Ma, and Kaizheng Wang. Spectral method and regularized mle are both optimal for top-k ranking. *Annals of statistics*, 47(4):2204, 2019.

[17] Bruce Choppin. A fully conditional estimation procedure for rasch model parameters. 1982.

[18] Paulo Cortez and Alice Maria Gonçalves Silva. Using data mining to predict secondary school student performance. 2008.

[19] Mark A Davenport, Yaniv Plan, Ewout Van Den Berg, and Mary Wootters. 1-bit matrix completion. *Information and Inference: A Journal of the IMA*, 3(3):189–223, 2014.

[20] Persi Diaconis and Laurent Saloff-Coste. Comparison theorems for reversible markov chains. *The Annals of Applied Probability*, 3(3):696–730, 1993.

[21] Susan Embretson. A general latent trait model for response processes. *Psychometrika*, 49(2):175–186, 1984.

[22] Gerhard H Fischer. On the existence and uniqueness of maximum-likelihood estimates in the rasch model. *Psychometrika*, 46(1):59–77, 1981.

[23] Mary Garner Jr. An eigenvector method for estimating item parameters of the dichotomous and polytomous rasch models. *Journal of Applied Measurement*, 3(2):107–128, 2002.

[24] Malay Ghosh. Inconsistent maximum likelihood estimators for the rasch model. *Statistics & Probability Letters*, 23(2):165–170, 1995.

[25] Shelby J Haberman. Maximum likelihood estimates in exponential response models. *The annals of statistics*, 5(5):815–841, 1977.

[26] Shelby J Haberman. Joint and conditional maximum likelihood estimation for the rasch model for binary responses. *ETS Research Report Series*, 2004(1):i–63, 2004.

[27] Ronald K Hambleton, Hariharan Swaminathan, and H Jane Rogers. *Fundamentals of item response theory*, volume 2. Sage, 1991.

[28] F Maxwell Harper and Joseph A Konstan. The movielens datasets: History and context. *Acm transactions on interactive intelligent systems (tiis)*, 5(4):1–19, 2015.

[29] David R Hunter and Kenneth Lange. A tutorial on mm algorithms. *The American Statistician*, 58(1):30–37, 2004.

[30] Sadiq Hussain, Neama Abdulaziz Dahan, Fadl Mutaher Ba-Alwib, and Najoua Ribata. Educational data mining and analysis of students' academic performance using weka. *Indonesian Journal of Electrical Engineering and Computer Science*, 9(2):447–459, 2018.

[31] Denis Kotkov, Alan Medlar, Alexandr Maslov, Umesh Raj Satyal, Mats Neovius, and Dorota Glowacka. The tag genome dataset for books. In *Proceedings of the 2022 ACM SIGIR Conference on Human Information Interaction and Retrieval (CHIIR'22)*, volume 5, 2022.

[32] John P Lalor, Hao Wu, and Hong Yu. Learning latent parameters without human response patterns: Item response theory with artificial crowds. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*, 2019.

[33] Andrew Lan, Mung Chiang, and Christoph Studer. An estimation and analysis framework for the rasch model. In *International Conference on Machine Learning*, pages 2883–2891. PMLR, 2018.

[34] Erich L Lehmann and George Casella. *Theory of point estimation*. Springer Science & Business Media, 2006.

[35] John Michael Linacre. Conditional maximum likelihood estimation.

[36] John Michael Linacre. Estimation methods: Jmle, prox, wmle, cmle, pmle, amle.

[37] Frederic M Lord. *Applications of item response theory to practical testing problems*. Routledge, 2012.

[38] R Duncan Luce. *Individual choice behavior: A theoretical analysis*. Courier Corporation, 2012.

[39] Geoff N Masters. A rasch model for partial credit scoring. *Psychometrika*, 47(2):149–174, 1982.

[40] Lucas Maystre. Choix: Inference algorithms for models based on luce's choice axiom. `https://github.com/lucasmaystre/choix`, 2015.

[41] Roderick P McDonald. *Test theory: A unified treatment*. psychology press, 2013.

[42] Prathiba Natesan, Ratna Nandakumar, Tom Minka, and Jonathan D Rubright. Bayesian prior choice in irt estimation using mcmc and variational bayes. *Frontiers in psychology*, 7:1422, 2016.

[43] Sahand Negahban, Sewoong Oh, and Devavrat Shah. Rank centrality: Ranking from pairwise comparisons. *Operations Research*, 65(1):266–287, 2017.

[44] James R Norris and James Robert Norris. *Markov chains*. Number 2. Cambridge university press, 1998.

[45] Georg Rasch. Studies in mathematical psychology: I. probabilistic models for some intelligence and attainment tests. 1960.

[46] Alexander Robitzsch. A comprehensive simulation study of estimation methods for the rasch model. *Stats*, 4(4):814–836, 2021.

[47] Pedro Rodriguez, Joe Barrow, Alexander Miserlis Hoyle, John P Lalor, Robin Jia, and Jordan Boyd-Graber. Evaluation examples are not equally informative: How should that change nlp leaderboards? In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4486–4503, 2021.

[48] Roseanna W Saaty. The analytic hierarchy process—what it is and how it is used. *Mathematical modelling*, 9(3-5):161–176, 1987.

[49] Ryan Sanchez. GIRTH: G. Item Response Theory , 11 2021.

[50] Carolin Schellhorn and Rajneesh Sharma. Using the rasch model to rank firms by managerial ability. *Managerial Finance*, 2013.

[51] Wim J Van der Linden and RK Hambleton. Handbook of item response theory. *Taylor & Francis Group. Citado na pág*, 1(7):8, 1997.

[52] Roman Vershynin. *High-dimensional probability: An introduction with applications in data science*, volume 47. Cambridge university press, 2018.

[53] Pauli Virtanen, Ralf Gommers, Travis E. Oliphant, Matt Haberland, Tyler Reddy, David Cournapeau, Evgeni Burovski, Pearu Peterson, Warren Weckesser, Jonathan Bright, Stéfan J. van der Walt, Matthew Brett, Joshua Wilson, K. Jarrod Millman, Nikolay Mayorov, Andrew R. J. Nelson, Eric Jones, Robert Kern, Eric Larson, C J Carey, İlhan Polat, Yu Feng, Eric W. Moore, Jake VanderPlas, Denis Laxalde, Josef Perktold, Robert Cimrman, Ian Henriksen, E. A. Quintero, Charles R. Harris, Anne M. Archibald, Antônio H. Ribeiro, Fabian Pedregosa, Paul van Mulbregt, and SciPy 1.0 Contributors. SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods*, 17:261–272, 2020.

https://doi.org/10.52202/068431-2813

[54] Jacob Whitehill, Ting-fan Wu, Jacob Bergsma, Javier Movellan, and Paul Ruvolo. Whose vote should count more: Optimal integration of labels from labelers of unknown expertise. *Advances in neural information processing systems*, 22, 2009.

[55] Benjamin D Wright and Geofferey N Masters. *Rating scale analysis*. MESA press, 1982.

[56] Cai-Nicolas Ziegler, Sean M McNee, Joseph A Konstan, and Georg Lausen. Improving recommendation lists through topic diversification. In *Proceedings of the 14th international conference on World Wide Web*, pages 22–32, 2005.

[57] Aeilko H Zwinderman. Pairwise parameter estimation in rasch models. *Applied Psychological Measurement*, 19(4):369–375, 1995.